19

Construction of Phylogenetic Trees

Walter M. Fitch and Emanuel Margoliash

Reprinted from Science, January 20, 1967, Vol. 155, No. 3760, pages 279-284

Reprinted from Science, January 20, 1967, Vol. 155, No. 3760, pages 279-284

Construction of Phylogenetic Trees

A method based on mutation distances as estimated from cytochrome *c* sequences is of general applicability.

Walter M. Fitch and Emanuel Margoliash

Biochemists have attempted to use quantitative estimates of variance between substances obtained from different species to construct phylogenetic trees. Examples of this approach include studies of the degree of interspecific hybridization of DNA (*1*), the degree of cross reactivity of antisera to purified proteins (*2*), the number of differences in the peptides from enzymic digests of purified homologous proteins, both as estimated by paper electrophoresis-chromatography or column chromatography and as estimated from the amino acid compositions of the proteins (*3*), and the number of amino acid replacements between homologous proteins whose complete primary structures had been determined (*4*). These methods have not been completely satisfactory because (i) the portion of the genome examined was often very restricted, (ii) the variable measured did not reflect with sufficient accuracy the mutation distance between the genes examined, and (iii) no adequate mathematical treatment for data from large numbers of species was available. In this paper we suggest several improvements under categories (ii) and (iii) and, using cytochrome *c*, for which much precise information on amino acid sequences is available, construct a tree which, despite our examining but a single gene, is remarkably like the classical phylogenetic tree that has been obtained from purely biological data (*5*). We also show that the analytical method employed has general applicability, as exemplified by the derivation of appropriate relationships among ethnic groups from data on their physical characteristics (*6, 7*).

Dr. Fitch is an assistant professor of physiological chemistry at the University of Wisconsin Medical School in Madison. Dr. Margoliash is head of the Protein Section in the Department of Molecular Biology, Abbott Laboratories, North Chicago, Illinois.

Determining the Mutation Distance

The *mutation distance* between two cytochromes is defined here as the minimal number of nucleotides that would need to be altered in order for the gene for one cytochrome to code for the other. This distance is determined by a computer making a pairwise comparison of homologous amino acids (*8*). For each pair a *mutation value* is taken from Table 1 which gives the minimum number of nucleotide changes required to convert the coding from one amino acid to the other. The table is derived from Fig. 2 of Fitch (*9*) except that, as a result of the work of Weigert and Garen (*10*) and Brenner, Stretton, and Kaplan (*11*), the uridyl-adenosylpurine trinucleotide is now treated as a chain-terminating codon. This change of codon meaning, although it does not affect the method of calculation, does cause the mutation values for amino acid pairs involving glutamine with cysteine, phenylalanine, tyrosine, serine, and tryptophan to become 1 greater than in the table previously published (*12*). Also, misprints involving the leucine-glycine and valine-cysteine pairs have been corrected. To maintain homology, deletions, all of which occur near the ends of the chains, are represented by X's. The amino- and carboxyl-terminal sequences in which deletions occur are shown in Table 2. Thus all cytochromes are regarded as being 110 amino acids long. If the homologous pairing includes an X, no mutation value is assigned.

For each possible pairing of cytochromes, the 110 mutation values found are summed to obtain the minimal mutation distance. For purposes of calculation, these mutation distances are proportionally adjusted to compensate for variable numbers of pairs of residue positions in which at least one member contains an X. For example, the number of X-containing amino acid pairs occurring between the *Saccharomyces* and *Candida* cytochromes *c* is 1, whereas that between two mammalian cytochromes *c* is 6. Thus the known mutation distance of the former pairing is multiplied by 110/109 whereas that of the latter is multiplied by 110/104. The results for 20 known cytochromes *c*, rounded off to the nearest whole number, are shown in the lower left half of Table 3.

The basic approach to the construction of the tree is illustrated in Fig. 1, which shows three hypothetical proteins, A, B, and C, and their mutation distances. There are two fundamental problems: (i) Which pair does one join together first? (ii) What are the lengths of legs *a*, *b*, and *c*?

As a first approximation, one solves problem (i) simply by choosing the pair with the smallest mutation distance, which in this case is A and B, with a distance of 24. Hence A and B are shown connected at the lower apex in Fig. 1. To solve the second problem, one notes that the distance from A to C, 28, is 4 less than the distance from B to C. Hence there must have been at least 4 more countable mutations in the descent of B from the lower apex than in the descent of A. Thus if $a + b = 24$ and $b - a = 4$, then $a = 10$, $b = 14$, and therefore $c = 18$. Note that an exact solution is obtained from which a reconstruction of the mutation distances precisely matches the input data.

Table 1. Mutation values for amino acid pairs. Each value is the minimum number of nucleotides that would need to be changed in order to convert a codon for one amino acid into a codon for another. The table is symmetrical about the diagonal of zeros. Letters across the top represent the amino acids in the same order as in the first column and conform to the single-letter code of Keil, Prusik, and Sŏrm (*21*).

	A	C	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	W	Y
Aspartic acid	0	2	2	2	1	1	2	1	3	1	1	2	2	2	2	3	2	1	2	1
Cysteine	2	0	2	1	3	2	3	2	3	2	1	2	3	1	1	1	2	2	2	1
Threonine	2	2	0	2	2	2	1	1	1	1	2	1	2	1	1	2	2	2	1	2
Phenylalanine	2	1	2	0	3	2	3	2	2	2	1	2	3	2	1	2	1	1	1	2
Glutamic acid	1	3	2	3	0	2	1	1	2	2	2	2	1	2	2	2	2	1	3	1
Histidine	1	2	2	2	2	0	2	2	3	1	1	1	1	1	2	3	1	2	2	2
Lysine	2	3	1	3	1	2	0	2	1	1	2	2	1	1	2	2	2	2	2	2
Alanine	1	2	1	2	1	2	2	0	2	2	2	1	2	2	1	2	2	1	2	1
Methionine	3	3	1	2	2	3	1	2	0	2	3	2	2	1	2	2	1	1	1	2
Asparagine	1	2	1	2	2	1	1	2	2	0	1	2	2	2	1	3	2	2	1	2
Tyrosine	1	1	2	1	2	1	2	2	3	1	0	2	2	2	1	2	2	2	2	2
Proline	2	2	1	2	2	1	2	1	2	2	2	0	1	1	1	2	1	2	2	2
Glutamine	2	3	2	3	1	1	1	2	2	2	2	1	0	1	2	2	1	2	3	2
Arginine	2	1	1	2	2	1	1	2	1	2	2	1	1	0	1	1	1	2	2	1
Serine	2	1	1	1	2	2	2	1	2	1	1	1	2	1	0	1	1	2	1	1
Tryptophan	3	1	2	2	2	3	2	2	2	3	2	2	2	1	1	0	1	2	3	1
Leucine	2	2	2	1	2	1	2	2	1	2	2	1	1	1	1	1	0	1	1	2
Valine	1	2	2	1	1	2	2	1	1	2	2	2	2	2	2	2	1	0	1	1
Isoleucine	2	2	1	1	3	2	2	2	1	1	2	2	3	2	1	3	1	1	0	2
Glycine	1	1	2	2	1	2	2	1	2	2	2	2	2	1	1	1	2	1	2	0

Table 2. Areas of cytochromes *c* involving deletions. The first seven and the last four amino acids of the cytochromes *c* for the 20 species studied are shown. Deletions are represented by X's. Sequences are reported in the single-letter code of Keil *et al.* (*21*), a key to which is provided in Table 1.

Amino terminal positions 1–7	Organism	Carboxyl terminal positions 107–110
PLPFGQY	*Candida*	LXSI
XEGFILY	*Saccharomyces*	LXCG
XXYFSLY	*Neurospora*	LELX
XXYVPLY	Moth	SEXI
XXYVPLY	Screwworm fly	LSEI
XXXXXXY	Tuna	LESX
XXXXXXY	All other vertebrates	No deletions

When information from more than three proteins is utilized, the basic procedure is the same, except that initially each protein is assigned to its own subset. One then simply joins two subsets to create a single, more comprehensive, subset. This process is repeated according to the rules set forth below until all proteins are members of a single subset. A phylogenetic tree is but a graphical representation of the order in which the subsets were joined.

In the present case, we start with 20 subsets, each subset consisting of a single cytochrome *c* amino acid sequence. To determine which two subsets should be joined, all possible pairwise combinations of subsets are in turn assigned to sets *A* and *B*, with all remaining subsets in each case assigned to set *C*. In each alternative test all proteins are thus a part of one of the three sets. The three sets are treated exactly as in the preceding example, except that now the mutation distances used are averages determined from every possible pairing of proteins, one from each of the two sets whose average mutation distance is being calculated.

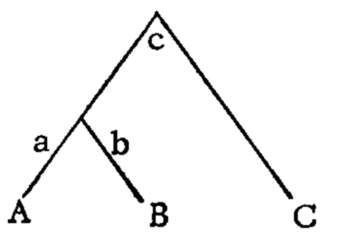
One arbitrarily accepts, from among all the possible pairings examined, that assignment of protein subsets to sets *A*, *B*, and *C* which provides the lowest average mutation distance from *A* to *B*. The leg lengths are then calculated and recorded. Henceforth the proteins of *A* and *B* so joined are treated as a single subset, and the entire procedure described in the preceding paragraph is repeated. Thus the number of subsets, originally equal to the number of proteins (N), is reduced by 1 with each cycle. In this fashion, after N-1 joinings of subsets, the initial phylogenetic tree will have been produced. Because average mutation distances are now being used, the solutions obtained are very unlikely to permit an exact reconstruction of the input data.

Testing Alternative Trees

Because of the arbitrary nature of the rule by which proteins are assigned to sets A and B, the initial tree will not necessarily represent the best use of the information. To examine reasonable alternatives, one simply constructs another tree by assigning an alternative pair of protein subsets to sets A and B whenever the mutation distance between the two subsets is not greater by some arbitrary amount than that between the members of the initial pair used in constructing the initial phylogenetic tree (*13*). The tree that is less satisfactory on the basis of criteria set forth below is discarded, and other alternatives are tested.

The best of 40 phylogenetic trees so far examined is presented in Fig. 2. Each juncture is located on the ordinate at a point representing the average of all distances between the juncture and the species descendant from it. The mutation distance to any one descendant may be more or less than the ordinate value.

By summing distances over the tree, it is possible to reconstruct values (upper right half of Table 3) comparable to the original input mutation distances (lower left half of Table 3).



Mutation Distances

	B	C
A	24	28
B		32

Fig. 1. Calculation of observed mutation distances. The upper apex represents a hypothetical ancestral organism that divided into two descending lines, one of which subsequently also divided. Thus we have three present-day species, A, B, and C. The number of observable mutations that have occurred in a particular gene since the A and B lines of descent diverged are represented respectively by *a* and *b*. The number of mutations that separate the lower apex and C is represented by *c*. The sums of $a + b$, $a + c$, and $b + c$, then, are the mutation distances of the three species as currently observed.

The 20 species are indicated in the last column; the identifying numbers in the first column and the top row of the table may be used as coordinates. Thus the tabulated values interrelating the human and horse cytochromes at coordinates (1,4) and (4,1) are mutation distances of 17 and 15 respectively, the former being the input datum, the latter having been obtained from the tree by reconstruction. If the absolute difference between two such mutation distances $| (i,j) - (j,i) |$ is multiplied by 100 and divided by (i,j), the result is the percentage of change from the input data. If such values are squared and the squares are summed over all values of $i < j$, the resultant sum (Σ) may be used to obtain the percent "standard deviation" (*4*) of the reconstructed values from the input mutation distances. The number of mutation distances summed is $N(N-1)/2$, or 190 for our case. If this number is reduced by 1, divided into the sum Σ, and the square root taken, the result is the percent "standard deviation." Since the standard deviation is a larger number than the standard error, the probable error, or the average deviation, the percent "standard deviation" is used here, it being less likely to create overconfidence in the significance of a result (*4*).

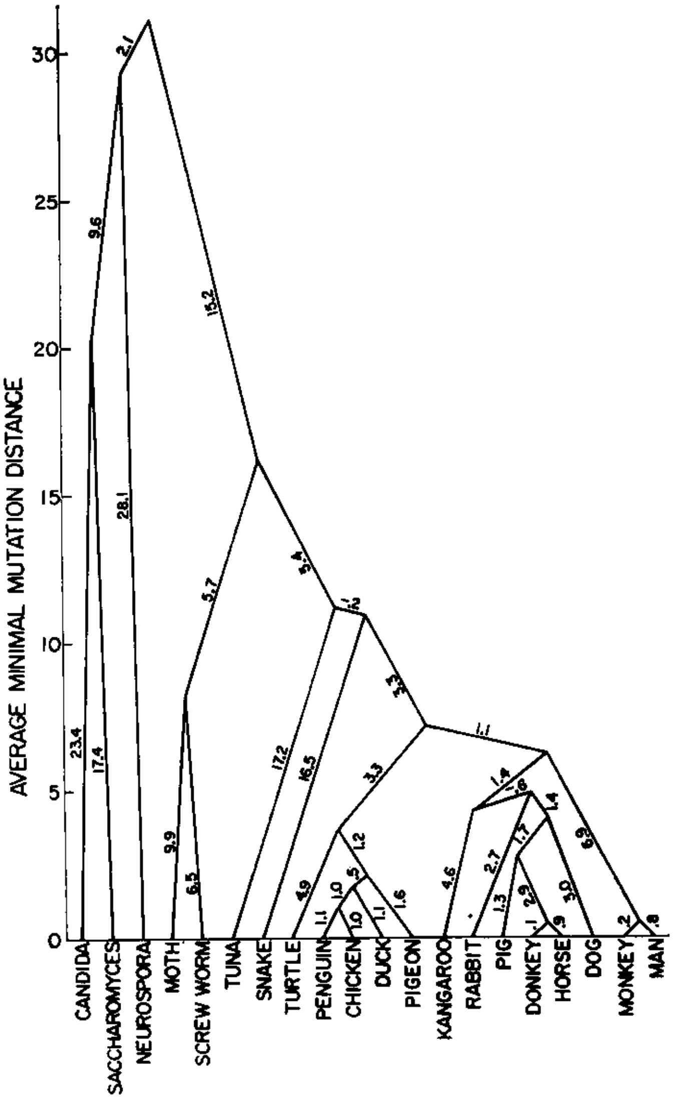
The Statistically Optimal Tree

In testing phylogenetic alternatives, one is seeking to minimize the percent "standard deviation." The scheme shown in Fig. 2 has a percent "standard deviation" of 8.7, the lowest of the 40 alternatives so far tested. The percent "standard deviation" for the initial tree was 12.3.

In addition to using a gene product to discover evolutionary relationships among several species, one can similarly delineate evolutionary relationships among different genes. Our procedure constructs, from the amino acid sequences of human alpha, beta, gamma, and delta hemoglobin chains and whale myoglobin (*15*), the gene phylogeny

Table 3. Minimum numbers of mutations required to interrelate pairs of cytochromes *c*. Values in the lower left half of the table are mutation distances as determined from the amino acid sequences and, prior to rounding off, were used to derive Fig. 2. Values in the upper right half of the table are reconstructed distances found by summing the leg lengths in Fig. 2. The references cited in the last column are to studies of the amino acid sequences of the cytochromes *c* of the indicated species.

Protein	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1		1	13	15	15	13	11	14	15	15	16	16	17	29	29	30	33	64	62	68	Man (*22*)
2	1		12	15	14	12	11	13	15	14	15	15	16	28	29	29	32	63	61	67	Monkey (*Macacus mulatta*) (*23*)
3	13	12		9	8	6	7	8	13	13	13	14	15	26	27	27	30	61	59	65	Dog (*24*)
4	17	16	10		1	5	10	11	15	15	16	16	17	29	29	30	33	64	62	68	Horse (*25*)
5	16	15	8	1		4	9	10	14	14	15	15	16	28	28	29	32	63	61	67	Donkey (*26*)
6	13	12	4	5	4		7	8	13	12	13	13	14	26	27	27	30	61	59	65	Pig (*27*)
7	12	11	6	11	10	6		7	11	11	12	12	13	24	25	25	29	60	57	63	Rabbit (*30*)
8	12	13	7	11	12	7	7		13	13	14	14	15	27	27	28	31	62	60	66	Kangaroo (*Canopus canguru*) (*28*)
9	17	16	12	16	15	13	10	14		3	3	3	8	26	27	27	30	61	59	65	Pekin duck (*29*)
10	16	15	12	16	15	13	8	14	3		4	4	8	26	27	27	30	61	59	65	Pigeon (*29*)
11	18	17	14	16	15	13	11	15	3	4		2	9	27	27	28	31	62	60	66	Chicken (*17*)
12	18	17	14	17	16	14	11	13	3	4	2		9	27	27	28	31	62	60	66	King penguin (*Aptenodytes patagonica*) (*29*)
13	19	18	13	16	15	13	11	14	7	8	8	8		28	29	29	32	63	61	67	Snapping turtle (*Chelydra serpentina*) (*31*)
14	20	21	30	32	31	30	25	30	24	24	28	28	30		33	34	37	68	66	72	Rattlesnake (*Crotalus adamanteus*) (*32*)
15	31	32	29	27	26	25	26	27	26	27	26	27	27	38		35	38	69	67	73	Tuna (*33*)
16	33	32	24	24	25	26	23	26	25	26	26	28	30	40	34		16	59	56	63	Screwworm fly (*Haematobia irritans*) (*29*)
17	36	35	28	33	32	31	29	31	29	30	31	30	33	41	41	16		62	60	66	Moth (*Samia cynthia*) (*34*)
18	63	62	64	64	64	64	62	66	61	59	61	62	65	61	72	58	59		56	62	*Neurospora* (*crassa*) (*35*)
19	56	57	61	60	59	59	59	58	62	62	62	61	64	61	66	63	60	57		41	*Saccharomyces* (*oviformis*) iso-1 (*36*)
20	66	65	66	68	67	67	67	68	66	66	66	65	67	69	69	65	61	61	41		*Candida* (*krusei*) (*37*)



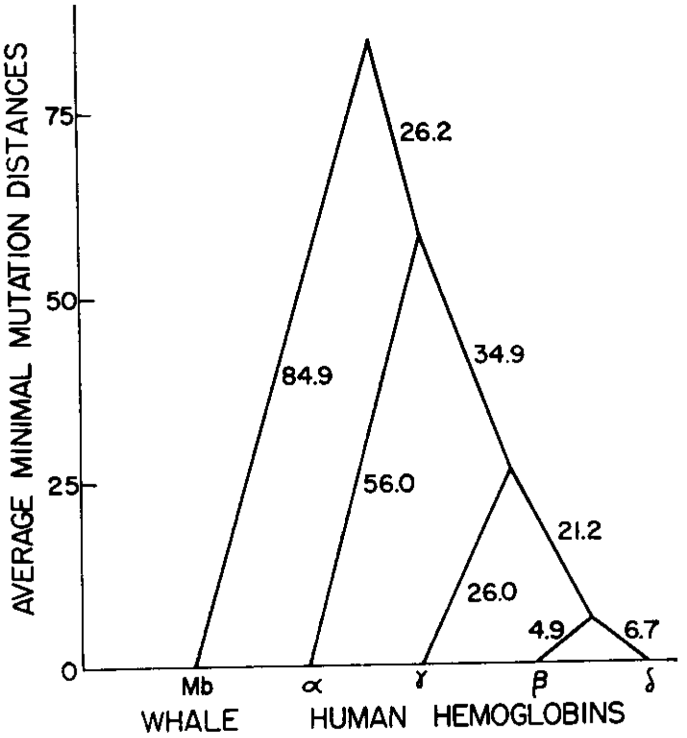


Fig. 2 (left). Phylogeny as reconstructed from observable mutations in the cytochrome *c* gene. Each number on the figure is the corrected mutation distance (see text) along the line of descent as determined from the best computer fit so far found. Each apex is placed at an ordinate value representing the average of the sums of all mutations in the lines of descent from that apex.

Fig. 3 (right above). A gene phylogeny as reconstructed from observable mutations in several heme-containing globins. See Fig. 2 for details. The percent "standard deviation" (*7*) for this tree is 1.33.

Table 4. Descent of the mammalian cytochromes. Changes in amino acids are shown in large capitals, with subscripts to indicate the number of mutations that had to occur to produce the indicated change. In general, unchanging amino acids are not repeated, but occasionally it has been necessary to relist an unchanged amino acid because a mutation appearing in one line of descent did not apply to other lines listed further down the page. Such unchanged amino acids are shown in small capitals. The lines of descent are shown on either side of the table. The last two columns give the sum of the mutations indicated in that row and the corresponding value from Fig. 2. The following rules were used in formulating each amino acid position of the ancestral sequences: Choose the amino acid so that the changes in the codon during descent require (i) the smallest overall number of mutations; (ii) the fewest segments containing multiple mutations (that is, two lines with one mutation each are preferred to one line with two mutations); (iii) the fewest sequential mutations (that is, one mutation in each of two lines following a branch point is preferred to one mutation before and one after the branch point); (iv) the fewest back mutations; (v) the fewest kinds of amino acids. Rule (i), where applicable, took priority over all others and rule (ii) took priority over the remainder. It was not found necessary to choose among the last three rules. The ancestral mammalian cytochrome *c* sequence shown was derived from the amino acid sequences of all 20 cytochromes *c*.

																		Mutuations (No.)	
Amino acid No.	17	18	21	39	41	50	52	53	56	64	66	68	89	94	95	98	109	Listed in this table	From leg lengths in Fig. 2
Ancestral mammal	V	Q	L	H	U	P	O	S	A	E	Y	A	L	I	G	L	N		
Ancestral primate	W_1	M_2	S_1	.	.	.	.	.	L_1	.	.	.	V_1	.	.	.	.	6	6.9
Monkey	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0.2
Man	.	.	.	.	.	.	.	.	.	W_1	.	.	.	.	.	.	.	1	.8
	V	Q	L	.	.	.	F_1	.	A	E	.	.	L	.	Y_1	.	.	2	1.4
Kangaroo	.	.	.	N_1	W_1	.	.	E_1	.	W_1	.	.	.	.	.	.	.	4	4.6
	.	.	.	H	U	.	.	S	.	E	.	.	.	.	.	.	.	0	—.6
Rabbit	.	.	.	.	.	V_2	.	.	.	.	.	.	.	.	A_1	.	.	3	2.7
	.	.	.	.	.	P	.	.	.	.	.	G_1	.	.	Y	.	.	1	1.4
Dog	.	.	.	.	.	.	.	.	.	.	.	.	.	E_1	.	.	I_4	2	3.0
Ancestral ungulate	.	.	.	.	.	.	.	.	.	.	.	.	.	I	.	Q_2	N	2	1.7
Pig	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1.3
Ancestral perissodactyl	.	.	.	.	.	.	.	.	.	.	I_3	.	.	.	E_2	.	.	4	2.9
Donkey	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0.1
Horse	.	.	.	.	.	.	.	E_1	.	.	.	.	.	.	.	.	.	1	.9

shown in Fig. 3. The overall result is as Ingram had previously indicated (*15*). A cautionary note may be derived from this. A wildly incorrect result could easily be obtained if the presence of multiple, homologous genes were not recognized and a phylogeny were constructed from sequences which were coded for, say, half by genes for alpha hemoglobin chains and half by genes for beta hemoglobin chains. This results from the speciation having occurred more recently than the gene duplication which permitted the separate evolution of the alpha and beta genes.

The method described can also be used to develop treelike relationships by employing data which are very different in character from mutation distances. For example, the physical characteristics of human beings have been used to construct a tree relating several ethnic groups (Fig. 4; *6*).

Although we are examining the product of but a single gene, and a rather small one at that, the phylogenetic scheme in Fig. 2 is remarkably like that constructed in accord with classical zoological comparisons (*5*). There are only three noticeable deviations, discussed below, and these may well be changed as more species are added to the list. Of even greater value would be sequences from other genes, since special environmental effects may easily cause the convergence of one or several genes in phylogenetically disparate organisms. Hemoglobin amino acid sequences may soon be available in great enough numbers to prove useful in this respect.

Almost all the alternative phylogenetic schemes tested involved rearrangements within the groups birds (*16, 17*) and nonprimate mammals (*14, 18, 19*). With respect to the birds, it will be noticed that the penguin is closely associated with the chicken, whereas one might have expected that all the "birds of flight" (Neognathae) would be more closely related to each other than to the penguin (Impennae). This discrepancy is probably related to the very small numbers of mutations involved. In this regard, it is interesting to note that on the basis of a micro-complement-fixation technique using antisera to several purified enzymes, Wilson *et al.* (*2*) found that the duck is more closely related to the chicken than is the pigeon. This agrees with our findings.

In the second group, the kangaroo is shown closely associated with the nonprimate mammals, whereas most zoologists would maintain that the placental mammals, including the primates, are more closely related to each other than to the marsupials.

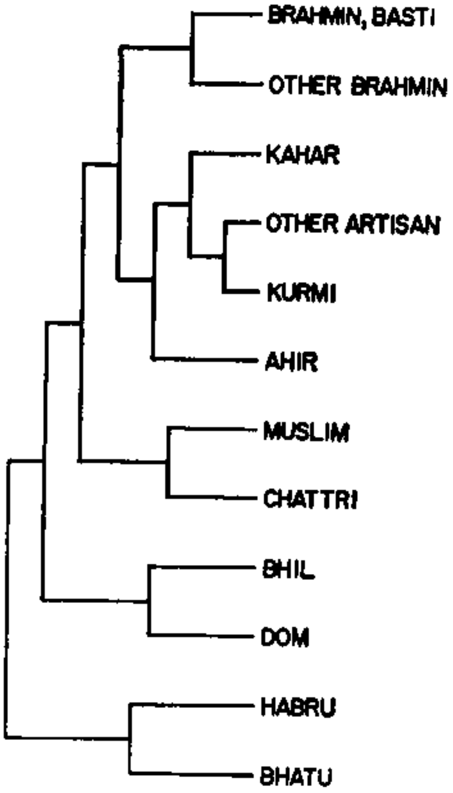


Fig. 4. Relations among various tribes and castes of India. The data used to construct this scheme are the D^2 values given by Rao (*6*). This figure is in principle like Fig. 2 except that, to prevent misinterpretation of the physical significance of the numbers one obtains, branching is shown as a rather uniform step function which preserves the relationships but obliterates the quantitative distances of the ordinate.

A third anomaly is that the turtle appears more closely associated with the birds than to its fellow reptile the rattlesnake. Although it is true that the snake is involved in seven of the nine instances where the reconstructed values differ from the input mutation distances by more than 4 mutations, this cannot account for the anomaly, which in fact results from the close similarity of the turtle's cytochrome *c* amino acid sequence to those of the birds.

Thus the phylogenetic tree in Fig. 2 is imperfect. Nevertheless, considering that only one gene product was analyzed and that no choices were made other than those dictated by the statistical analysis, the results are very promising, and a phylogeny based upon a quantitative determination of those very events which permit speciation, namely mutations, must ultimately be capable of providing the most accurate phylogenetic trees.

Elapsed Time and Evolutionary Change

It should be pointed out that the ordinate of Fig. 2 represents the minimum number of mutations observable. Since multiple mutations in a single codon are not likely to produce mutation values as large as the actual number of mutations sustained, Fig. 2 is greatly foreshortened with respect to the actual number of mutations (*20*). The possibility of obtaining an ordinate scale denoted as actual mutations by applying a correction factor, using the relative frequencies of codons observed to have sustained one, two, and three nucleotide changes, must await reliable statistical information on the relative probabilities that given amino acid substitutions will permit the progeny to compete successfully in their environment. Any meaningful correction of this sort is precluded at present by the lack of such statistical information, but its importance may be emphasized by noting that such a correction would yield an ordinate in Fig. 2 in which equal numbers of mutations would correspond to equal intervals of time, as long as the rate at which mutations are fixed, averaged for many lines of descent over very long periods of evolutionary history, does not vary appreciably (*20*).

It should be noted that the method does not assume any particular value for the rate at which mutations have accumulated during evolution. Indeed, from any phylogenetic ancestor, today's descendants are equidistant with respect to time but not, as computations show, equidistant genetically. Thus the method indicates those lines in which the gene has undergone the more rapid changes. For example, from the point at which the primates separate from the other mammals, there are, on the average, 7.5 mutations in the descent of the former and 5.8 in that of the latter, indicating that the change in the cytochrome *c* gene has been much more rapid in the descent of the primates than in that of the other mammals.

The method allows negative mutation distances, and a few were observed in some of the discarded phylogenetic schemes. Their absence from the best-fitting scheme would indicate that there were no significant evolutionary reversals in this gene.

One highly desirable goal is the reconstruction of the ancestral cytochrome *c* amino acid sequences. The procedure, though not difficult, is dependent upon the phylogenetic tree on which these sequence data are arranged. Given the present scheme (see Fig. 2) ,one can reconstruct the ancestral proteins. A reconstruction of the ancestral amino acid sequences for the mammalian portion of tree is shown in Table 4. One can then ask such a question as "What are the mutations required to account for the difference between the cytochromes *c* of the ancestral primate and of the ancestral mammal?" The data in Table 4 clearly identify the mutations as occurring in positions 17, 18, 21, 56, and 89. In a similar manner, the monkey and human lines are distinguished by a single mutation in the human line which resulted in the substitution of isoleucine for threonine at position 64.

There is presently no detectable relationship between the primary structures of cytochrome *c* and those of hemoglobins (*12*). Nevertheless, the reconstruction and comparison of the ancestral amino acid sequences may reveal a homology that cannot be detected in present-day proteins. The employment of such ancestral sequences may be generally useful for detecting common ancestry not otherwise observable.

Note added in proof. Since this article was accepted our attention has been called to several earlier papers which present some of the important concepts discussed here. Sokal and his collaborators (*38*) have for several years been studying various ways of producing treelike relationships from quantitative taxonomic information. In an interesting application of this type of technique, using the amino acid sequences of fibrinopeptides from several ungulates, R. F. Doolittle and B. Blombäck (*39*) constructed such a tree and specifically indicated how knowledge of the genetic code would be useful for more precise constructions.

Jukes (*40*, fig. 3) has presented the Ingram scheme of the hemoglobin gene duplications and placed upon the various legs estimates of the numbers of nucleotide substitutions. His figure is not essentially different from Fig. 3 of this article.

References and Notes

1. B. J. McCarthy and E. T. Bolton, *Proc. Natl. Acad. Sci. U.S.* **50**, 156 (1963).
2. A. C. Wilson, N. O. Kaplan, L. Levine, A. Pesce, M. Reichlin, W. S. Allison, *Fed. Proc.* **23**, 1258 (1964); C. A. Williams, Jr., in *Peptides of Biological Fluids*, H. Peeters, Ed. (Elsevier, New York, 1965), p. 62; M. Goodman, *ibid.*, p. 70; A. S. Hafleigh and C. A. Williams, Jr., *Science* **151**, 1530 (1966).
3. R. L. Hill, J. Buettner-Janusch, V. Buettner-Janusch, *Proc. Natl. Acad. Sci. U.S.* **50**, 885 (1963); R. L. Hill and J. Buettner-Janusch, *Fed. Proc.* **23**, 1236 (1964).
4. E. Margoliash, *Proc. Natl. Acad. Sci. U.S.* **50**, 672 (1963); E. L. Smith and E. Margoliash, *Fed. Proc.* **23**, 1243 (1964); E. Margoliash and E. L. Smith, in *Evolving Genes and Proteins*, V. Bryson and H. Vogel, Eds. (Academic Press, New York, 1965), p. 221.
5. A. S. Romer, *Vertebrate Paleontology* (Univ. of Chicago Press, Chicago, ed. 2, 1945).
6. Our procedure may be compared with the "cluster analysis" approach as formulated by A. W. F. Edwards and L. L. Cavalli-Sforza [*Biometrics* **21**, 362 (1965)]; their approach is, in one sense, the reverse of that we have used, since cluster analysis starts with all the elements as members of the same subset and proceeds to subdivide that subset into successively smaller but more numerous subsets until each element is the sole member of its own subset. In terms of Fig. 2, Edwards and Cavalli-Sforza constructed their tree from the top down, whereas we built ours from the bottom up. Edwards and Cavalli-Sforza report testing their method on C. R. Rao's data [*Advanced Statistical Methods in Biometric Research* (Wiley, New York, 1952)] on physical characteristics of 12 Indian castes and tribes. Rao had used these data to postulate relationships among the castes and tribes. Although the nature of these data is quite different from that of ours, the formal mathematical problems are very much alike, and we have used the D^2 values of Rao, as did Edwards and Cavalli-Sforza, to find the best tree. Edwards and Cavalli-Sforza's tree has a percent "standard deviation" (*7*) of 32.6. Our result, shown in Fig. 4, has a percent "standard deviation" of 29.2 and, except that it possesses greater detail, conforms to the conclusions drawn by Rao.
7. The quotation marks are placed around "standard deviation" because the data used in its formulation here are not statistically independent as is generally required. This is evident in that only 20 amino acid sequences determine the 190 mutation distances utilized.
8. The homology may be found by aligning the cysteine residues which bind the heme. Excellent examples of this may be seen in Fig. 10 of E. Margoliash and A. Schejter, *Advan. Protein Chem.* **20**, 114 (1965).
9. W. M. Fitch, *J. Mol. Biol.* **16**, 1 (1966).
10. M. G. Weigert and A. Garen, *Nature* **206**, 992 (1965).
11. S. Brenner, A. O. W. Stretton, S. Kaplan, *ibid.*, p. 994.
12. W. M. Fitch, *J. Mol. Biol.* **16**, 9 (1966).
13. It will be recognized that once the first tree is calculated, the number of computations required for alternatives becomes greatly reduced. For example, if instead of the tree shown in Fig. 2 one wishes to test a tree which differs only in the order in which the chicken, duck, and penguin are joined, the only legs in need of recalculation are those five descending to these birds from the avian apex.
14. The cow (*18*) and sheep (*19*) cytochromes *c* are identical with that of the pig (*27*).
15. V. M. Ingram, *The Hemoglobins in Genetics and Evolution* (Columbia Univ. Press, New York, 1963); A. B. Edmundson, *Nature* **205**, 883 (1965).
16. The cytochrome *c* of the turkey (*29*) is identical with that of the chicken (*16*).
17. S. K. Chan and E. Margoliash, *J. Biol. Chem.* **241**, 507 (1966).
18. K. T. Yasunobu, T. Nakashima, H. Higo, H. Matsubara, A. Benson, *Biochim. Biophys. Acta* **78**, 791 (1963).
19. S. K. Chan, S. B. Needleman, J. W. Stewart, E. Margoliash, unpublished results.
20. This is analogous to the relationship between numbers of amino acid replacements and the evolutionary time scale discussed by E. Margoliash and E. L. Smith in *Evolving Genes and Proteins*, V. Bryson and H. Vogel, Eds. (Academic Press, New York, 1965), p. 221.
21. B. Keil, Z. Prusik, F. Sorm, *Biochim. Biophys. Acta* **78**, (1963).
22. H. Matsubara and E. L. Smith, *J. Biol. Chem.* **238**, 2732 (1963).
23. J. A. Rothfus and E. L. Smith, *ibid.* **240**, 4277 (1965).
24. M. A. McDowall and E. L. Smith, *ibid.* p. 4635.
25. E. Margoliash, E. L. Smith, G. Kreil, H. Tuppy, *Nature* **192**, 1125 (1961).
26. O. F. Walasek and E. Margoliash, unpublished results.
27. J. W. Stewart and E. Margoliash, *Can. J. Biochem.* **43**, 1187 (1965).
28. C. Nolan and E. Margoliash, *J. Biol. Chem.* **241**, 1049 (1966).
29. S. K. Chan, I. Tulloss, E. Margoliash, unpublished results.
30. S. B. Needleman and E. Margoliash, *J. Biol. Chem.* **241**, 853 (1966).
31. S. K. Chan, I. Tulloss, E. Margoliash, *Biochemistry* **5**, 2586 (1966).
32. O. P. Bahl and E. L. Smith, *J. Biol. Chem.* **240**, 3585 (1965).
33. G. Kreil, *Z. Physiol. Chem.* **334**, 154 (1963).
34. S. K. Chan and E. Margoliash, *J. Biol. Chem.* **241**, 335 (1966).
35. J. Heller and E. L. Smith, *Proc. Natl. Acad. Sci. U.S.* **54**, 1621 (1965).
36. Y. Yaoi, K. Titani, K. Narita, *J. Biochem. Tokyo* **59**, 247 (1966).
37. K. Narita and K. Titani, *Proc. Japan Acad.* **41**, 831 (1965).
38. R. R. Sokal, *Syst. Zool.* **10**, 70 (1961); F. J. Rohlf and R. R. Sokal, *Univ. Kansas Sci. Bull.* **45**, 3 (1965); J. H. Camin and R. R. Sokal, *Evolution* **19**, 311 (1965).
39. R. F. Doolittle and B. Blombäck, *Nature* **202**, 147 (1964).
40. T. H. Jukes, *Advan. Biol. Med. Phys.* **9**, 1 (1963).
41. This project received support from grants from NIH (NB-04565) and NSF (GB-4017) to W.M.F. We thank Peter Guetter and Daniel Brick for valuable technical assistance.